# Datasheet for Dataset

Konstantinos Tzortzakis

## Motivation for Dataset Creation

### Why was the dataset created?

The need to retain customers is huge in telecommunication companies because of the competitive landscape. As a result, Telco Customer Churn dataset was created to provide this telecommunications company the ability to predict customer churn. Customer churn occurs when customers stop doing business with a company, also known as customer attrition. Telecom companies use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one.

### What (other) tasks could the dataset be used for?

Every appropriate task for this dataset comes under the end goal of understanding customers' behavior. Therefore, it can be used to monitor customers' reaction to new services or contracts as well as different needs that a specific company might have. It should not be used for any task that does not involve the relationship between this specific company and those 7,043 customers. Finally, it can also be used as a learning tool in order to explore this type of models and learn more about the subject.

### Has the dataset been used for any tasks already?

The dataset has only been used by different people with different approaches in order to predict customer churn.

https://www.kaggle.com/blastchar/telco-customer-churn

https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/

### Who funded the creation of the dataset?

Unknown

## Dataset Composition

### What are the instances?

Each instance is a customer of this company and general information about him/her, demographics, services that have signed up for and account information (e.g. Contract, Monthly Charges)

### Are relationships between instances made explicit in the data

There are no known relationships between instances except for the fact that they are all customers of this telecom company.

### How many instances of each type are there?

The data consists of 7,043 customers and 21 features for every customer.

### What data does each instance consist of?

Each instance consists of demographic information with regards to customers:

- customer id
- gender (male, female)
- SeniorCitizen (0,1)
- Partner (yes, no)
- Dependents (yes, no)

Services that each customer has signed up for:

- PhoneService (yes, no)
- MultipleLines (yes, no, no phone service
- InternetService (Fiber optic, DSL, no)
- OnlineSecurity (yes, no, no internet service)
- OnlineBackup (yes, no, no internet service)
- DeviceProtection (yes, no, no internet service)
- TechSupport (yes, no, no internet service)
- StreamingTV (yes, no, no internet service)
- StreamingMovies (yes, no, no internet service)

And account information:

- Contract (one year, two year, month to month)
- PaperlessBilling (yes, no)
- PaymentMethod (electronic check, mailed check, bank transfer(automatic), credit card(automatic))
- MonthlyCharges (number) in $
- TotalCharges (number) in $
- Tenure (min: 1, max:72) in months

- Churn (yes, no): customers who left within the last month

**Is everything included or does the data rely on external resources?**

Everything is included.

**Are there recommended data splits or evaluation measures?**

There are not recommended data splits or evaluation measures based on the extensive search concerning the use of this dataset.

**What experiments were initially run on this dataset?**

There are 191 kernels available in:
https://www.kaggle.com/blastchar/telco-customer-churn

These users have tried to find the best prediction using different methods such as Logistic Regre ssion, Random Forest and ANN and well-known libraries such as numpy, pandas, matplotlib.pyp lot and seaborn.

Data analysis is also provided by ibm.com in order to demonstrate the use of Watson Analytics as a tool for analysis.

https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/

## Data Collection Process

Unfortunately, there is no information concerning the data collection process. We only know that the owner is ibm.com. I searched all 191 kernels in Kaggle.com and no one refers to the

source of the data. To justify this I provide information from the discussion addressing this particular concern:
https://www.kaggle.com/blastchar/telco-customer-churn/discussion/82495

where the dataset creator (updated 1 year ago, for the first time):
https://www.kaggle.com/blastchar/telco-customer-churn

mention that the original dataset, as referred, is from IBM Watson Analytics community. This community only provides the dataset and some tools as mentioned before to demonstrate the use of Watson Analytics.

**Over what time-frame was the data collected?**

The only time-frame information we have is that this dataset includes customers who left within the last month

## Data Preprocessing

**What preprocessing/cleaning was done?**

No preprocessing/cleaning

## Dataset Distribution

**How is the dataset distributed?**

Website publicly available through ibm.com

https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA_Fn-UseC_-Telco-Customer-Churn.csv?cm_mc_uid=42350349240115511137586&cm_mc_sid_50200000=48695861551226681996&cm_mc_sid_52640000=68226791551226682000

**When will the dataset be released/first distributed?**

The dataset was first released and distributed by McKinley Stacker IV on April 15, 2015. Then introduced to kaggle.com by the user " BlastChar".

**What license (if any) is it distributed under?**

Based on the extensive search there is not a distribution license.

**Are there any fees or access/export restrictions? Any other comments?**

There are neither restrictions nor fees and the dataset is available for anyone who is interested.

## Dataset Maintenance

**Who is supporting/hosting/maintaining the dataset?**

IBM - McKinley Stacker IV
Kaggle – "BlastChar"

**Will the dataset be updated?**

Unknown

**If the dataset becomes obsolete how will this be communicated?**

Unknown

**Is there a repository to link to any/all papers/systems that use this dataset?**

There is not a repository to link all systems that use this dataset except for Kaggle's users:

https://www.kaggle.com/blastchar/telco-customer-churn/kernels

**If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?**

N/A

## Legal and Ethical Considerations

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?**

It is related to people (telecom's customers) but it is unknown if they have agreed for its use. As I mentioned before there is no information about the data collection process.

**If it relates to people, were there any ethical review applications/reviews/approvals?**

Unknown

**If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?**

Unknown

**If it relates to people, could this dataset expose people to harm or legal action?** (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

To mitigate the potential for harm, the data owner does not reveal the names of the customers. Hence, these people are not exposed at all.

**If it relates to people, does it unfairly advantage or dis- advantage a particular social group?**

Fortunately, the sample is fairly distributed between males and females, including single people and people with partners as well as parents and senior citizens. Therefore, this dataset does not unfairly advantage or dis–advantage any particular social group.

**If it relates to people, were they provided with privacy guarantees?**

Unknown

**Does the dataset comply with the EU General Data Protection Regulation (GDPR)?**

Unknown

**Does the dataset contain information that might be considered sensitive or confidential?**

Although the dataset contains certain information about these people, it cannot be considered sensitive or confidential. They are protected by providing only their customer id that no one is able to relate to the actual names of the customers.